Multiple Imputation of Missing Multivariate Atmospheric Chemistry Time Series Data from Denali National Park

By

Chanachai Charoonsophonsak, B.S. Mathematics Statistics Concentration

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Statistics

University of Alaska Fairbanks

May 2020

APPROVED:

Dr. Scott Goddard, Committee Chair
Dr. Ronald Barry, Committee Member
Dr. Julie McIntyre, Committee Member
Dr. Margaret Short, Committee Member
Dr. Leah Berman, Chair
Department of Mathematics and Statistics
Dr. Kinchel C. Doerner, Dean
College of Natural Science and Mathematics
Dr. Michael Castellini, *Dean of the Graduate School*

Abstract

This paper explores a technique where we impute missing values for an incomplete dataset via multiple imputation. Incomplete data is one of the most common issues in data analysis and often occurs when measuring chemical and environmental data. The dataset that we used in the model consists of 26 atmospheric particulates or elements that were measured semiweekly in Denali National Park from 1988 to 2015. The collection days were alternating between three and four days apart from 3/2/88 - 9/30/00 and being consistently collected every three days apart from 10/3/00 - 12/29/15. For this reason, the data were initially partitioned into two in case the separation between collection days would have an impact. With further analysis, we concluded that the misalignments between the two datasets had very little or no impact on our analysis and therefore combined the two. After running five Markov chains of 1000 iterations we concluded that the model stayed consistent between the five chains. We found out that in order to get a better understanding of how well the imputed values did, more exploratory analysis on the imputed datasets would be required.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work was supported in part by the high-performance computing and data storage resources operated by the Research Computing Systems Group at the University of Alaska Fairbanks Geophysical Institute. Special thanks to Cathy Cahill for the dataset that she provided. I would also like to thank the members of my committee. Lastly, I would like to thank my advisor Dr. Goddard for the countless of hours of work he helped me on this project. This project would not have been completed without your help, thank you.

1 Introduction

The dataset that we obtained comes from Denali National Park in Alaska that monitors visibility impairment in the air. This data that we obtained starts on September 27, 1986 and ends on December 29, 2015. However, the dataset that we ended up using starts on March 2, 1988 and ends on December 29, 2015. The reason why we don't use any data prior to March 2, 1988 is due to the inconsistency of the number of days between collection days, which we will discuss in Section 2.

Incomplete data often occurs when measuring chemical and environmental data and this was definitely the case with the data we obtained. Incomplete data can be defined as missing data (or missing values). This is an issue, as normal analyzing techniques will not be applicable to incomplete data. Incomplete data can occur for a variety of reasons, such as an interruption of a scheduled collection time, which results in fully missing values.

Another type of incomplete data may be censored data. This type of incompleteness is caused by the limitations of the instruments. For example, an instrument may not be able to register a value due to it being below the detection limits. Depending on the instrument's setting, it will either report a value such as zero or be missing.

In either cases of censored or missing data it is a nuisance when trying to analyze the data. We resolve this issue by creating one or more complete datasets by filling in the missing data by means of either single imputation or multiple imputation.

Single imputation is a method in which a missing value is replaced by exactly one value. An example of single imputation may be applying a simple linear regression to estimate the missing value. The advantages of using single imputation are that calculations tend to be simple, they only need to be carried out once, and they are computationally inexpensive. Some disadvantages are that the imputed values may be misleading, i.e. it may result in poor predictions, or introduce bias, and perhaps most importantly do not reflect the sampling variability.

Multiple imputation, on the other hand, replaces each missing value with several values. Instead of creating a single complete dataset, multiple imputation creates n complete datasets. Each complete dataset is then analyzed using complete-data methods as if the imputed data were the real data. This method ignores the distinction between the data that was observed and the data that was imputed. The n completed datasets can then be combined to produce a final inference, such as taking the mean or median of the n completed datasets. The advantages of multiple imputation are that by imputing multiple values for each missing or censored value (rather than just one imputed value), we incorporate the uncertainty of what these values are, and this also tends to reduce bias. The disadvantages to multiple imputation are the increased complexity of models, which results in more difficult calculations, and this means the calculations are or can be computationally expensive.

In Section 2 we will discuss in more detail the dataset from Denali National Park and the decision to use a subset of the data we obtained. In Section 3, we will discuss the two models we implemented on the dataset. In Section 4, we will discuss the results we got from the second model that was implemented, and finally we will end with discussion and future work in Section 5.

2 Data

The data we obtained are airborne particulate samples that were collected semi-weekly in Denali National Park in Alaska by the National Park Service under Interagency Monitoring of Protected Visual Environment (IMPROVE). One of the issues when measuring chemical and environmental data is that incomplete data often occurs and this was certainly the case with the dataset. The data consist of 26 particulate constituents (or elements) obtained between September 1986 and December 2015, with a total of 3181 samples, which was mostly measured semi-weekly. The 26 particulate constituents in this dataset are used to determine $PM_{2.5}$ and PM_{10}, which ultimately is used to assess the visibility impairment in Denali National Park. This is why the dataset is often referred to as the Visibility dataset. According to the metadata file, the units used for the Visibility dataset are micrograms of the individual chemical species per cubic meters ($\mu g/m^3$).

There are missing values that occur in the Visibility dataset from time to time. From the status flags in the data, we determined that none of the missing values were caused due to being below the detection limit. That is, none of the missing data are caused by limitations of the instrument; all missing data were classified as fully missing. The cause of the missingness is uncertain. Possible reasons include: human negligence, instrument failure, record loss, and weather uncooperativeness. A possible example of weather uncooperativeness is when the instrument is frozen due to the negative temperatures. It may be that the values of a certain collection were purposefully deleted for some reason. However, we will assume that none of the values was deliberately deleted and there is no pattern of fully missing values between collections days. This is referred to as "missing completely at random" or MCAR.

In the dataset there are cases where an entry was missing, which means all 26 constituents or elements are missing for a particular collection day. And in other cases there may be a combination of some missing and non-missing elements for a given collection day. This means that there are a total of $(2^{26} - 1) = 67108863$ possible combinations of missing values to

consider. The Visibility dataset that we obtained started on September 27, 1986 and ended on December 29, 2015, however, no prior collection days were used prior to March 2, 1988. The reason for this decision is due to the inconsistent number of days between collection between September 1986 and March 1988. Between September 1986 and March 1988 there were 10 collection periods where the sample collected was 7 days apart and one collection period where it was 98 days apart.

Therefore we decided use only the portion of the dataset between March 2, 1988 and December 29, 2015. On further investigation of the dataset, we noticed that there was a change between the days of collection period on October 3, 2000. Between March 2, 1988 and September 30, 2000 the data was being collected every Wednesday and Sunday, alternating between collection days of three and four days apart. Between October 3, 2000 and December 29, 2015 we noticed that the data was being consistently collected every three days apart. Therefore we considered partitioning dataset into two datasets just in case the separation between collection days would have an impact. The number and percentage of fully missing values of the elements are summarized in tables in the next two sections. In Section 2.3 we explain the reason why we decided to use the combined dataset from March 2, 1988 to December 29, 2015 without partitioning.

2.1 Visibility Dataset 1

As previously stated, we noticed that between March 2, 1988 and September 30, 2000, the days between collection period was three and four days apart. This meant that in the first subset of the dataset, data was being collected every Wednesday and Sunday. Visibility Dataset 1, the first subset, had a total of 1314 samples. The element Molybdenum (Mo) has the highest percentage of fully missing values. Table 2.1 below lists the number of missing values, non-missing values, and the percentage missing in the Visibility Dataset 1.

Element	Fully Missing	Non-Missing	Percentage Missing
Al, As, Br, Ca, Cl, Cu, Fe	38	1276	2.89
Pb, Mg, Ni, P, K, Rb, Se	38	1276	2.89
Si, Na, Sr, S, Ti, V, Zn, Zr	38	1276	2.89
Cr, Mn	39	1275	2.97
Hf	44	1270	3.35
Mo	579	735	44.06

Table 2.1: Summary of missing data for Visibility Data 1

2.2 Visibility Dataset 2

Between October 3, 2000 and December 29, 2015, data was being consistently collected every three days apart. According to a paper on Air Quality Monitoring at Denali National Park & Park preserve, Alaska 2000 - 2003, on October 1, 2000 the IMPROVE schedule was changed to match the EPA national schedule of 1-in-3 (2012). This meant that between October 3, 2000 and December 29, 2015, Visibility dataset 2, the second subset of the dataset, had a total of 1856 samples. Molybdenum (Mo) was no longer being recorded.

Element	Fully Missing	Non-Missing	Percentage Missing
Al, As, Br, Ca, Cl, Cr, Cu	70	1786	3.77
Fe, Pb, Mg, Mn, Ni, P, K	70	1786	3.77
Rb, Se, Si, Na, Sr, S, Ti	70	1786	3.77
V, Zn, Zr	70	1786	3.77
Hf	669	1187	36.05
Mo	1856	0	100

Table 2.2: Summary of missing data for Visibility Data 2

2.3 Combined Visibility Data

We initially thought that it would be better to separate the dataset into two subsets, but upon looking at the summary we decided it would be better if we didn't subset the data. There were advantages that we saw that made sense not to subset and to use the "combined data". For example, when taking a look at Table 2.2, the element Mo is missing all samples in dataset 2 and therefore values of imputation would not be as accurate. By combining the data, we can leverage the information of dataset 1 and understand how Mo behaves over time prior to October 3, 2000 and this ultimately results in better imputation.

Since we saw a change in the collection days between dataset 1 and dataset 2 we decided we might be able to justify the action by looking at the graphs, summary, and autocorrelation function (ACF) of dataset 1, dataset 2, and the combined dataset for anything unusual. Instead of including all the graphs for the 26 elements we decided to focus on three randomly selected elements. The time series graphs, summaries, and ACF of the elements aluminum (Al), chromium (Cr), and potassium (K) are shown below.

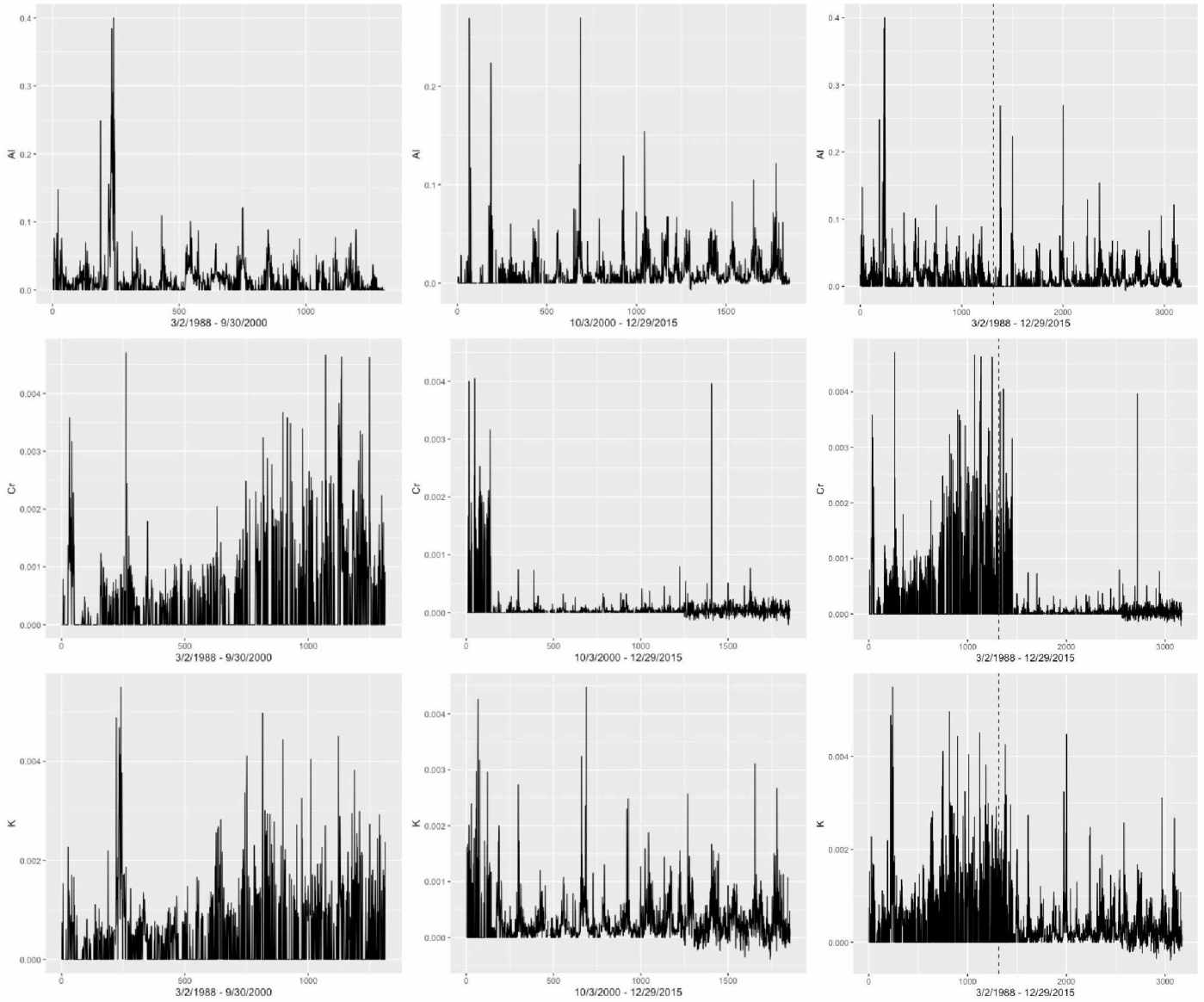

Figure 2.1: Time Series Graph

The first column of the graphs above are from dataset 1, the second column are from dataset 2, and the third column is the combined data. Row one represents aluminum (Al), row 2 chromium (Cr), and potassium (K) row 3. A dashed line was included for the combined data (third column), to indicate where dataset 1 ends and dataset 2 begins. The time series graphs are included to see whether if there is an obvious change from switching to the 1-in-3 schedule. It is important to note that there are negative values shown in the graphs in the

second column. After talking to experts who are in charge of the data, this is due to an algorithm change that occur in 2011 to take in account values that were indistinguishable from zero.

The only graph that may be concerning is Cr. The reason this may be concerning is because there was a big dip in the graph of the combined dataset (second row, third column). However, if we look at the graph and where the dashed line is relative to the where dip is, we can conclude that the big dip isn't caused from the change by switching to a 1-in-3 schedule.

The following table of summary statistics helps us understand what we're seeing from the plots above.

Aluminum (Al)	Min.	1st Qrt .	Median	Mean	3rd Qrt.	Max
Dataset 1	0.00	0.00	0.00632	0.01600	0.02060	0.40
Dataset 2	-0.00704	0.00	0.00445	0.01080	0.01390	0.270
Combined Dataset	-0.00704	0.00	0.00495	0.01290	0.0170	0.40

Chrominum (Cr)	Min.	1st Qrt.	Median	Mean	3rd Qrt.	Max
Dataset 1	0.00	0.00	0.00	0.00040	0.00062	0.00471
Dataset 2	-0.00021	0.00	0.00001	0.00008	0.00006	0.00450
Combined Dataset	-0.00021	0.00	0.00	0.00022	0.00009	0.00471

Potassium (K)	Min.	1st Qrt.	Median	Mean	3rd Qrt.	Max
Dataset 1	0.00	0.00	0.00	0.00047	0.00080	0.00550
Dataset 2	-0.00039	0.00005	0.0017	0.00030	0.00041	0.00448
Combined Dataset	-0.00039	0.00	0.00013	0.00037	0.00051	0.00550

Table 2.3: Summary of Datasets

From the summary above there are not any alarming values that stick out. We now look at the autocorrelation function (ACF), which is included in the next page.

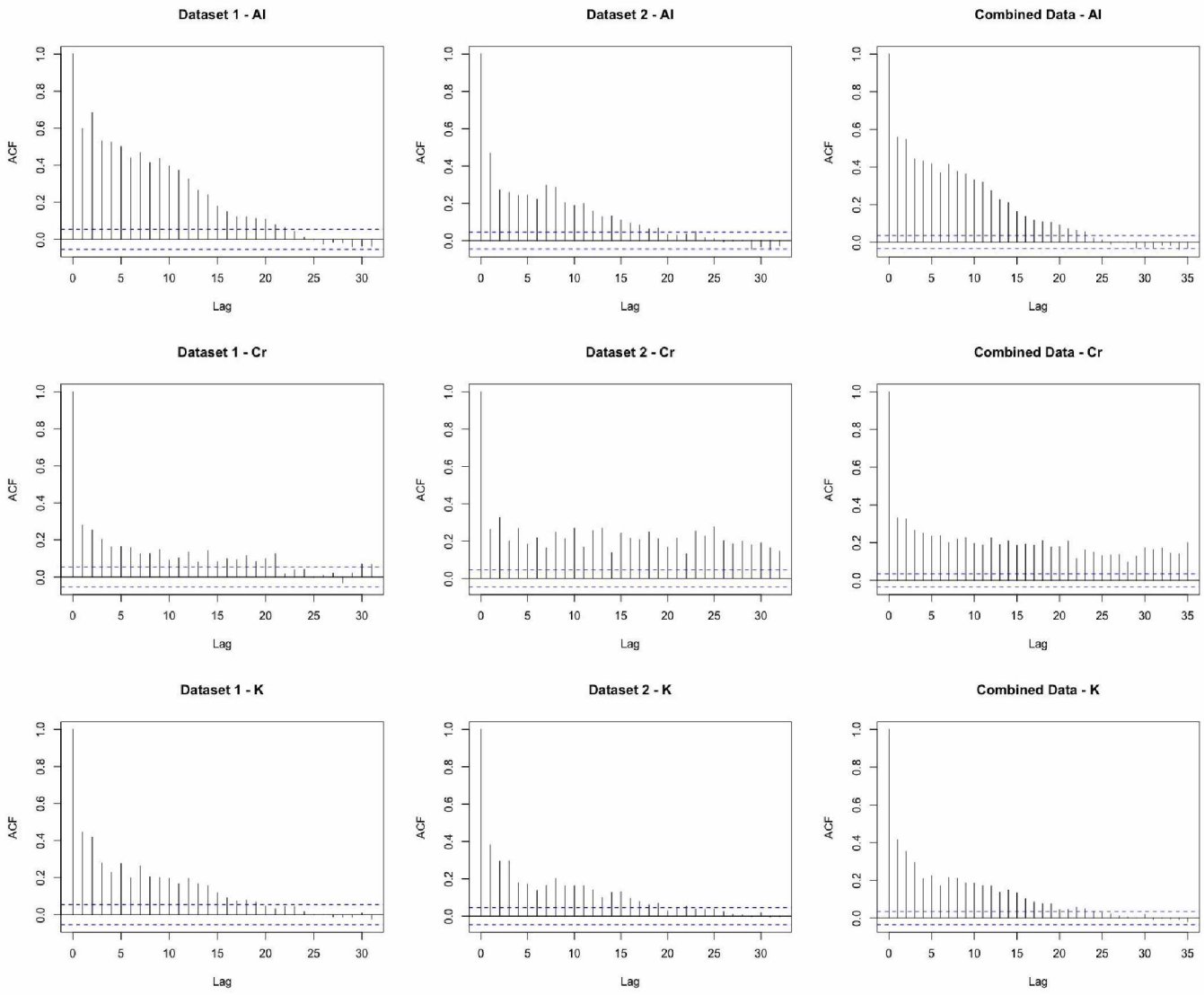

Figure 2.2: Graphs of ACF

From the graphs above, the ACF of each element almost looks identical in dataset 1, dataset 2 and the combined dataset. After looking looking at the time series graphs, statistics summaries, and ACF's we concluded that combining datasets 1 and 2 would not unduly affect our inferences. By combining the two datasets we now have a start date of March 2, 1988 and end date of December 29, 2015 with T = 3,170 total indices (i.e. collection days).

3 Methodology

The models used to impute the missing values come from a paper written by Hopke, Liu, and Rubin that proposed three models: The Cross-Sectional Model, The Multivariate Integrated Moving Average (IMA) Time Series Model, and The Multivariate IMA Seasonal Time Series Model (Hopke et al 2001). We implemented two of these models: The Cross-Sectional Model and The Multivariate Integrated Moving Average (IMA) Time Series Model. All three models require the usage of multiple imputation to handle the missingness in our dataset, and in order to achieve this a Gibbs Sampler was implemented using R.

The multiple imputation requires a Bayesian model in which the missing values are treated as parameters to be estimated. We assign prior distribution to these parameters, then use the data to specify a posterior distribution for the missing values. MCMC is used to obtain correlated samples from the posterior distribution, and these are what are used to find credible intervals for the missing values. A Gibbs sampler is a specific type of MCMC that can be used for certain types of statistical models including the ones that were used in this project. The details will be provided in the later section.

As previously stated, one of the downsides of multiple imputation is how computationally expensive it is. In order to increase the efficiency we ran our code on the supercomputer that is operated by the Research Computing Systems Group at the University of Alaska Fairbanks Geophysical Institute. For this project, we decided that it would be sufficient enough to run five Markov chains: that is, we ran the Gibbs Sampler five times, where each Gibbs Sampler was independent of each other. In order to achieve the five Markov chains we made five copies of the code we wrote in R and ran them as separate tasks on the supercomputer, in order for them to be independent. Since each iteration took around 10 minutes to run, we decided it would be sufficient if each Markov chain (that is, each Gibbs Sampler) ran for 1500 iterations. Due to the time constraint on the supercomputer, each chain could run a max of 300 iterations. In order to achieve the 1500 iterations we took the final iteration and

used it as the starting point for the next 300 iterations. This allowed the Gibbs Sampler to pick up where it ended as if we ran 1500 iterations simultaneously. Since we ran 1500 iterations per Markov chains this meant that we would end up with 7500 total completed datasets.

A possible downside of Gibbs Sampling that we need to take into consideration is that each iteration is dependent on the previous iteration, so it is important to have a burn-in. A burn-in is when a certain amount of the iterations at the beginning of the chain are thrown out to make the draws less dependent on the starting point of the chain. In this case, there was no prior knowledge of the parameters (that is, the missing data) so we used a vague prior. Two possible issues with using a vague prior are that samples are correlated and the start of the chain may be in the tail end of the posterior distribution. Since these are two issues to take in account it is important that a sufficient amount of burn-ins were used in order to be less influenced by the starting values. Thus, the decision was to toss the first 500 iterations (33% of the iterations) as burn-in and to use 1000 iterations as viable imputes. This meant that we now have five Markov chains with 1000 completed datasets in each chain.

3.1 Cross-Sectional Model

The Cross-Sectional Model is a simple multivariate normal model that is use to illustrate multiple imputation. The model incorporates cross-correlation between the types of elements being measured, but does not incorporate autocorrelation across time. The multivariate normal model will produce a joint distribution for $p = 26$ elements at any single time, and it assumes the observations across time are independent of each other. That is, $\boldsymbol{y_t} \overset{i.i.d}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Psi})$ where $\boldsymbol{\mu}$ is the p-dimensional mean vector, $\boldsymbol{\Psi}$ is the unknown $(p \times p)$ positive definite variance-covariance matrix, and $t = 1, ..., T$ with $T = 3170$ total samples from the data. In order to achieve full conditionals for the covariances that would be easy to draw from, the conjugate Inverse-Wishart prior was used $\boldsymbol{\Psi} \sim IW_p(\boldsymbol{W}, q)$ where $q = 1$ are the degrees of freedom, and $\boldsymbol{W}$ is a $(p \times p)$ identity matrix. In order to understand how the model works, it is best

to execute the steps required for one iteration of the Gibbs Sampler.

3.1.1 Gibbs Sampler Step 1 - Impute missing values

The algorithm that is used in the Gibbs Sampler is dependent on the missing and non-missing indexes at time t of the observed data, $\boldsymbol{y_{t,obs}}$. The $\boldsymbol{y_{t,miss}}$ and $\boldsymbol{y_{t,nonmiss}}$ are the vector of NA's or values that corresponds to missing and non-missing indices at time t. For example, suppose the observed data, $\boldsymbol{y_{t,obs}}$, was the following where $T = 20$ and $p = 5$.

t	Var 1	Var 2	Var 3	Var 4	Var 5
$t = 1$	NA	NA	1	NA	2
$t = 2$	3	NA	NA	4	5
		$\vdots$			
$t = 20$	6	7	NA	NA	8

The first step of the Gibbs Sampler is to first look at $t = 1$. At $t = 1$ the missing indexes are 1,2,4 and the non-missing indexes are 3,5.

t	Var 1	Var 2	Var 3	Var 4	Var 5
$t = 1$	NA	NA	1	NA	2
$t = 2$	3	NA	NA	4	5
		$\vdots$			
$t = 20$	6	7	NA	NA	8

These indexes at time t determines how $\boldsymbol{y_{t,miss}}$ is going to be drawn. As initial values of the Gibbs Sampler the vector of 1's $(1 \times p)$ was chosen for $\boldsymbol{\mu}$ and the identity matrix $(p \times p)$ was chosen for $\boldsymbol{\Psi}$. The vector of missing values, $\boldsymbol{y_{t,miss}}$ is drawn from the following conditional multivariate normal distribution.

$$\boldsymbol{y_{t,miss}}|\boldsymbol{\Theta} \sim MVN_p\left(\boldsymbol{\mu_{t,miss}} - (\boldsymbol{\Psi^{-1}_{miss}})^{-1}(\boldsymbol{\Psi^{-1}_{t,nonmiss:miss}})^{\top}\left(\boldsymbol{y_{t,nonmiss}} - \boldsymbol{\mu_{t,nonmiss}}\right), (\boldsymbol{\Psi^{-1}_{t,miss}})^{-1}\right)$$

where $\left(\boldsymbol{\Theta} = \boldsymbol{y_{t,nonmiss}}, \boldsymbol{\mu_{t,nonmiss}}, \boldsymbol{\Psi^{-1}_{t,obs:miss}}, \boldsymbol{\mu_{t,miss}}, \boldsymbol{\Psi^{-1}_{t,miss}}\right)$

Thus at $t = 1$, we must construct the vectors $\boldsymbol{y_{1,obs}}$, $\boldsymbol{y_{1,nonmiss}}$, and $\boldsymbol{y_{1,miss}}$ by using the missing and non-missing indexes.

$$\boldsymbol{y_{1,obs}} = \left(\begin{array}{|c|c|c|c|c|} \hline NA & NA & 1 & NA & 2 \\ \hline \end{array} \right)$$

$$\boldsymbol{y_{1,nonmiss}} = \Big(y_1[3], y_1[5]\Big)^{\top} = \begin{pmatrix} 1 & 2 \end{pmatrix}^{\top}$$

$$\boldsymbol{y_{1,miss}} = \Big(y_1[1], y_1[2], y_1[4]\Big)^{\top} = \begin{pmatrix} NA & NA & NA \end{pmatrix}^{\top}$$

Since each $\boldsymbol{\mu}$ will be unique depending on the missing and non-missing indexes at t, it makes sense to use the notation $\boldsymbol{\mu_{t,nonmiss}}$ and $\boldsymbol{\mu_{t,miss}}$. For example, $\boldsymbol{\mu_{1,nonmiss}}$ would be a subset of $\boldsymbol{\mu}$ using the non-missing indexes at $t = 1$.

$$\boldsymbol{\mu} = \left(\begin{array}{|c|c|c|c|c|} \hline 1 & 1 & 1 & 1 & 1 \\ \hline \end{array} \right)$$

Thus at $t = 1$, we have non-missing indexes of 3 and 5 and $\boldsymbol{\mu_{1,nonmiss}}$ (green color) is constructed as the following.

$$\boldsymbol{\mu_{1,nonmiss}} = \Big(\boldsymbol{\mu}[3], \boldsymbol{\mu}[5]\Big)^{\top} = \begin{pmatrix} 1 & 1 \end{pmatrix}^{\top}$$

Likewise we use the missing indexes at $t = 1$ to construct $\boldsymbol{\mu_{1,miss}}$ (red color) which is the following.

$$\boldsymbol{\mu_{1,miss}} = \Big(\boldsymbol{\mu}[1], \boldsymbol{\mu}[2], \boldsymbol{\mu}[4], \Big)^{\top} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^{\top}$$

Keep in mind that the formula for the density of the multivariate normal distribution requires $\boldsymbol{\Psi}^{-1}$, but since we chose the initial $\boldsymbol{\Psi}$ as the identity matrix then $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{-1}$. Likewise since each $\boldsymbol{\Psi}$ is unique depending on the missing and non-missing indices at t it would also make sense to use the notation $\boldsymbol{\Psi}^{-1}_{t,miss}$ and $\boldsymbol{\Psi}^{-1}_{t,nonmiss:miss}$.

Therefore in order to construct the matrices $\boldsymbol{\Psi}^{-1}_{1,miss}$ for $t = 1$ we would subset the matrix $\boldsymbol{\Psi}^{-1}$ using the missing indices as the subset.

$$\Psi^{-1} = \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right)$$

Therefore using the missing indexes of 1, 2 and 4 we construct $\boldsymbol{\Psi}^{-1}_{1,miss}$ (red color) as the following.

$$\boldsymbol{\Psi}^{-1}_{1,miss} = \begin{pmatrix} \Psi^{-1}[1,1] & \Psi^{-1}[1,2] & \Psi^{-1}[1,4] \\ \Psi^{-1}[2,1] & \Psi^{-1}[2,2] & \Psi^{-1}[2,4] \\ \Psi^{-1}[4,1] & \Psi^{-1}[4,2] & \Psi^{-1}[4,4] \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Likewise, we construct $\boldsymbol{\Psi}^{-1}_{1,nonmiss:miss}$ (green color) in a similar fashion except we uses a combination of the non-missing and missing indexes at time t, where the non-missing indexes are used as the row indices and the missing indexes are used as the column indices in the subset of $\boldsymbol{\Psi}^{-1}$.

$$\boldsymbol{\Psi}^{-1}_{1,nonmiss:miss} = \begin{pmatrix} \Psi^{-1}[3,1] & \Psi^{-1}[3,2] & \Psi^{-1}[3,4] \\ \Psi^{-1}[5,1] & \Psi^{-1}[5,2] & \Psi^{-1}[5,4] \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Once we have constructed all the vectors and matrices we can now draw $\boldsymbol{y_{1,miss}}$ from the following conditional distribution.

$$\boldsymbol{y_{1,miss}}|\boldsymbol{\Theta} \sim N_p\left(\begin{bmatrix}1\\1\\1\end{bmatrix} - \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}^{-1} \begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\end{bmatrix}^{\top} \left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix} \right), \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}^{-1} \right)$$

This process repeats for each $t = 1, 2, ..., T$ until all the missing values have been sampled. Once the data is filled with the imputed values for all T days, we can now reference the completed imputed data as $\boldsymbol{Y_{com}}$.

3.1.2 Gibbs Sampler Step 2 - Sample $\boldsymbol{\Psi}$, the covariance matrix for the p-variables

We first calculate $\boldsymbol{S_{yy^\top}}$. $\boldsymbol{S_{yy^\top}}$ which is $\boldsymbol{S_{yy^\top}} = (\boldsymbol{Y_{com}} - \bar{\boldsymbol{Y}})^\top(\boldsymbol{Y_{com}} - \bar{\boldsymbol{Y}})$, where $\bar{\boldsymbol{Y}} = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y_{t,com}}$, $\bar{\boldsymbol{Y}}$ is the mean of $\boldsymbol{y_{1,com}}, \boldsymbol{y_{2,com}}, ..., \boldsymbol{y_{T,com}}$. We now sample $\boldsymbol{\Psi}$ from its full conditional distribution:

$$\boldsymbol{\Psi}|\boldsymbol{Y_{com}} \sim IW_p(\boldsymbol{W} + \boldsymbol{S_{yy^\top}}, q + T),$$

3.1.3 Gibbs Sampler Step 3 - Sample $\boldsymbol{\mu}$, which has one component for each variable (element)

In this step we replace $\boldsymbol{\mu}$ by a new $\boldsymbol{\mu}$ that gets drawn from its following full conditional distribution,

$$\boldsymbol{\mu}|\boldsymbol{Y_{com}} \sim MVN_p(\bar{\boldsymbol{Y}}, \tfrac{1}{T}\boldsymbol{\Psi})$$

3.1.4 Gibbs Sampler Step 4 - Record sampled or imputed values for each iteration of the MCMC

Once Steps 1-3 have been completed, this is the end of an iteration of the Gibbs Sampler. Each time an iteration is completed we record $\boldsymbol{Y_{com}}$, $\boldsymbol{\Psi}$ and, $\boldsymbol{\mu}$.

3.2 Multivariate Integrated Moving Average (IMA) Time Series Model

Although the Cross-Sectional Model was easy to implement, it does not take into account a time series structure. In order to correct this, the first-order multivariate integrated moving average (IMA) time series model was implemented. Following Cleveland and Liu (1998), the IMA(1,1) process can be written as follows:

$$\boldsymbol{y_t} = \boldsymbol{x_t} + \boldsymbol{e_t} \text{ and } \boldsymbol{x_t} - \boldsymbol{x_{t-1}} = \boldsymbol{a_t} + \boldsymbol{a_{t-1}}$$

with the following properties:

$$\boldsymbol{e_t} \overset{i.i.d}{\sim} MVN_p(\mathbf{0}, \boldsymbol{\Psi}) \text{ and } \boldsymbol{a_t} \overset{i.i.d}{\sim} MVN_p(\mathbf{0}, \boldsymbol{\Omega})$$

where $\boldsymbol{x_t}$, $\boldsymbol{a_t}$, and $\boldsymbol{e_t}$ are p-dimensional vectors and $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$ are $(p \times p)$ positive definite variance-covariance matrices. The use of this representation of an IMA(1,1) process provides an easy method of finding the maximum likelihood estimates of the parameters and allows for a simple implementation of the Gibbs Sampler for multiple imputation (Dempster, Laird, and Rubin, 1997). Similar to the Cross-Sectional Model, in order to achieve full conditionals that would be easy to draw from, the conjugate semi conjugate Inverse-Wishart prior was used $\boldsymbol{\Psi} \sim IW_p(\boldsymbol{W_\Psi}, q_\Psi)$ and $\boldsymbol{\Omega} \sim IW_p(\boldsymbol{W_\Omega}, q_\Omega)$. Like the previous model, the hyperparameters $\boldsymbol{W_\Psi}$ and $\boldsymbol{W_\Omega}$ are $(p \times p)$ identity matrices and $q_\Psi = q_\Omega = 1$. The p-component vector $\boldsymbol{x_t}$ in this model plays the same role as $\boldsymbol{\mu}$ in the first model. Likewise, in order to understand the model it is best to execute the process it takes to run though an iteration of the Gibbs Sampler.

3.2.1 Gibbs Sampler Step 1 - Impute missing values

Suppose that we use the same data as the previous model where $p = 5$ and $T = 20$.

$t = 1$	Var 1	Var 2	Var 3	Var 4	Var 5
$t = 2$	NA	NA	1	NA	2
$t = 3$	3	NA	NA	4	5
		$\vdots$			
$t = 20$	6	7	NA	NA	8

The process of Step 1 is almost identical to the first model. The variable $\boldsymbol{\Psi}$ still plays the role of the variance-covariance matrix, except $\boldsymbol{x_t}$ plays the role of the mean vector. As initial values of the Gibbs Sampler, the matrix of 1's was chosen for $\boldsymbol{X}$ $(T \times p)$, the vector of 1's for $\boldsymbol{x_0}$ $(1 \times p)$, and the identity matrix $(p \times p)$ for both $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$. The vector of missing values, $\boldsymbol{y_{t,miss}}$ is drawn from the following multivariate normal distribution.

$$\boldsymbol{y_{t,miss}}|\boldsymbol{\Theta} \sim MVN_p\left(\boldsymbol{x_{t,miss}} - (\boldsymbol{\Psi^{-1}_{miss}})^{-1}(\boldsymbol{\Psi^{-1}_{t,nonmiss:miss}})^{\top}\left(\boldsymbol{y_{t,nonmiss}} - \boldsymbol{x_{t,nonmiss}}\right), (\boldsymbol{\Psi^{-1}_{t,miss}})^{-1}\right)$$

where $\left(\boldsymbol{\Theta} = \boldsymbol{y_{t,nonmiss}}, \boldsymbol{x_{t,nonmiss}}, \boldsymbol{\Psi^{-1}_{t,obs:miss}}, \boldsymbol{x_{t,miss}}, \boldsymbol{\Psi^{-1}_{t,miss}}\right)$

At $t = 1$ the process to construct the variables $\boldsymbol{y_{1,obs}}$, $\boldsymbol{y_{1,nonmiss}}$, $\boldsymbol{y_{1,miss}}$, and $\boldsymbol{\Psi}$ is the same as that of the Cross-Sectional model and the discussion will be skipped. The process to construct $\boldsymbol{x_{1,nonmiss}}$ and $\boldsymbol{x_{1,miss}}$ is similar except we are using $\boldsymbol{X}$ instead of $\boldsymbol{\mu}$.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ & & \vdots & & \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Likewise we are using the non-missing indexes at $t = 1$ to construct $\boldsymbol{x_{1,nonmiss}}$. Since the non-missing indexes at $t = 1$ are 3 and 5 we use these these values as the indices we use to subset $\boldsymbol{X}$ (green color).

$$\boldsymbol{x_{1,nonmiss}} = \Big(X[1,3], X[1,5]\Big)^\top = \begin{pmatrix} 1 & 1 \end{pmatrix}^\top$$

Similarly, we do the same thing for $\boldsymbol{x_{1,miss}}$ (red color) where the missing indexes are 1, 2, and 4 at $t = 1$.

$$\boldsymbol{x_{1,miss}} = \Big(X[1,1], X[1,2], X[1,4]\Big)^\top = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^\top$$

The construction of $\boldsymbol{\Psi}$ is identical to the previous model and therefore will be skipped.

3.2.2 Gibbs Sampler Step 2 - Sample $\boldsymbol{\Psi}$, the covariance matrix for the p-variable

We first calculate

$$\boldsymbol{S_{ee^\top}} = \sum_{i=1}^{T} (\boldsymbol{y_{t,com}} - \boldsymbol{x_t})(\boldsymbol{y_{t,com}} - \boldsymbol{x_t})^\top = (\boldsymbol{Y_{com}} - \boldsymbol{X})^\top (\boldsymbol{Y_{com}} - \boldsymbol{X})$$

We now sample $\boldsymbol{\Psi}$ from its following conditional distribution.

$$\boldsymbol{\Psi}|\boldsymbol{Y_{com}} \sim IW_p(\boldsymbol{W_\Psi}, +\boldsymbol{S_{ee^\top}}, q_\Psi + T)$$

3.2.3 Gibbs Sampler Step 3 - Sample $\boldsymbol{X}$, which has T components for each variable (element)

The next step in the Gibbs Sampler is to draw the matrix $\boldsymbol{X}$. In order to increase the efficiency of the algorithm we can use an orthogonal transformation. This allows us to draw the entire sequence of length T_p by drawing T independent components of length p, $\boldsymbol{G} = \boldsymbol{U_x^\top X}$,

where $\boldsymbol{U_x}$ is a $(T \times T)$ is the matrix of the eigenvectors that come from the spectral decomposition of $\boldsymbol{C}$, $\boldsymbol{C} = \boldsymbol{U_x \Lambda_x U_x^\top}$ and Λ_x is a $(T \times T)$ diagonal matrix of eigenvalues. The matrix $\boldsymbol{C}$ is constructed by recursively solving for the variance of $\boldsymbol{x_t = x_{t-1} + a_t + a_{t-1}}$. The matrix $\boldsymbol{C}$ is a $(T \times T)$ symmetric matrix with diagonals $\boldsymbol{C_{ij} = C_{ji}} = 2 + 4(i-1)$ for $i = j$ and $\boldsymbol{C_{ij}} = 3 + 4(i-1)$ for $j > i$ in the off-diagonals. In the case where we have $T = 20$ the $\boldsymbol{C}$ matrix is the following.

$$C = \begin{pmatrix} 2 & 3 & 3 & \cdots & 3 \\ 3 & 6 & 7 & \cdots & 7 \\ 3 & 7 & 10 & \cdots & 11 \\ \vdots & & & \ddots & \vdots \\ 3 & 7 & 11 & \cdots & 78 \end{pmatrix}$$

The whole matrix $\boldsymbol{G}$ gets drawn from the following full conditional.

$$\boldsymbol{G|Y_{com}, x_0} \sim MVN_{T_p}(\boldsymbol{F^{-1}H, F^{-1}})$$

where

$$\boldsymbol{F} = \boldsymbol{(\Lambda_x \otimes \Omega)^{-1} + (I_T \otimes \Psi)^{-1}}$$

$$\boldsymbol{H} = \text{vec}\big(\boldsymbol{(1_T^\top \otimes x_0)^\top U_x}\big)\boldsymbol{(\Lambda_x \otimes \Omega)^{-1}} + \text{vec}\big(\boldsymbol{Y_{com}^\top U_x}\big)\boldsymbol{(I_T \otimes \Psi)^{-1}}$$

and $\boldsymbol{1_T}$ is the vector of 1's of length T and $\boldsymbol{I_T}$ is a $(T \times T)$ identity matrix. Here $\otimes$ is the Kronecker product. Once the full matrix $\boldsymbol{G}$ is drawn, we can easily compute $\boldsymbol{X}$ as $\boldsymbol{U_x G}$ and the initial $\boldsymbol{X}$ gets updated.

3.2.4 Gibbs Sampler Step 4 - Sample $\boldsymbol{x_0}$, which has one componenet for each variable (element)

We sample $\boldsymbol{x_0}$ from the following full conditional distribution.

$$\boldsymbol{x_0}|\boldsymbol{X} \sim MVN_p\Big((1^\top \boldsymbol{C^{-1}} 1)^{-1} 1^\top \boldsymbol{C^{-1}} \boldsymbol{X}, (1^\top \boldsymbol{C^{-1}} 1)^{-1} \boldsymbol{\Omega}\Big)$$

3.2.5 Gibbs Sampler Step 5 - Sample $\boldsymbol{\Omega}$, the covariance matrix in charge of autocorrelation

We sample $\boldsymbol{\Omega}$ following full conditional distribution.

$$\boldsymbol{\Omega}|\boldsymbol{G}, \boldsymbol{x_0} \sim IW_p(\boldsymbol{W_\Omega} + (\boldsymbol{G} - \boldsymbol{U_x^\top} \otimes \boldsymbol{x_0})^\top \boldsymbol{\Lambda_x^{-1}} (\boldsymbol{G} - \boldsymbol{U_x^\top} \otimes \boldsymbol{x_0}), q_\Omega + T)$$

3.2.6 Gibbs Sampler Step 6 - Record sampled or imputed values for each iteration of the MCMC

Once Steps 1-5 have been completed, this is the end of an iteration of the Gibbs Sampler. Each iteration through we record $\boldsymbol{Y_{com}}$, $\boldsymbol{\Psi}$, $\boldsymbol{X}$, $\boldsymbol{x_0}$, and $\boldsymbol{\Omega}$. By the end this will result in n copies of $\boldsymbol{Y_{com}}$, $\boldsymbol{\Psi}$, $\boldsymbol{X}$, $\boldsymbol{x_0}$, and $\boldsymbol{\Omega}$ for every n iterations run though the Gibbs Sampler.

4 Results

Since we ran 5 Markov chains with 1000 iterations each we ended up with five sets of 1000 imputed datasets. This results in a complete dataset with 82,420,000 values per Markov chain and 412,100,000 total values among the chains. In order to get a sense of how the modeling worked we decided to choose an element to examine. We decided to look at the element Molybdenum (Mo) since every entry after October 3, 2000 was missing. We decided to look at trace plots of the five Markov chains of a random index after October 3 of Mo to see if it converged. The traces plots below are for a randomly selected collection day of Mo where $t = 1573$.

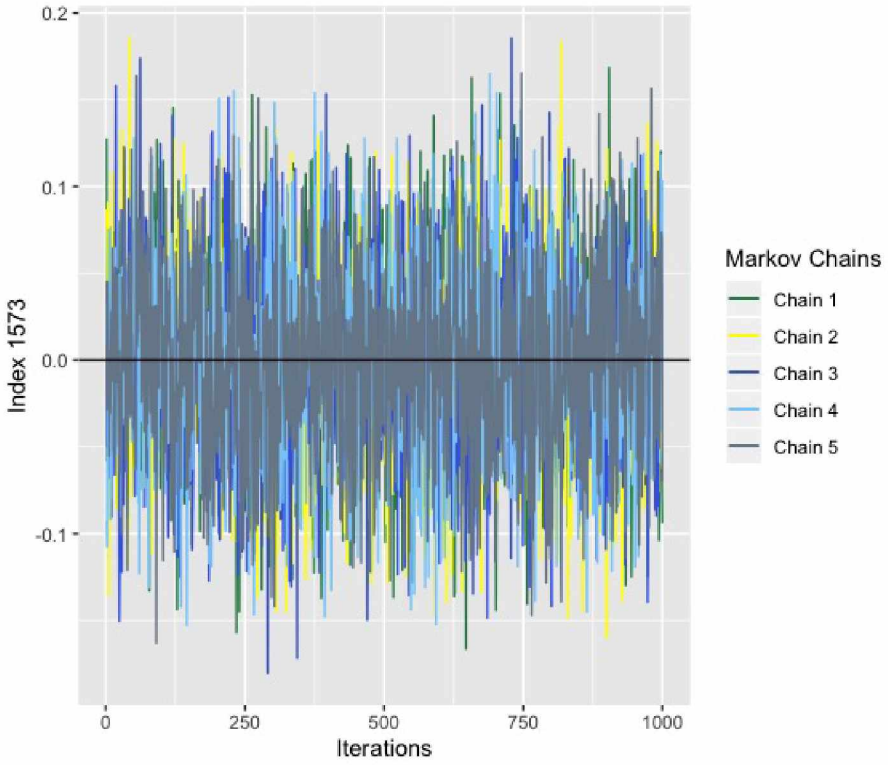

Figure 4.1: Trace Plots for Mo

The trace plots for the five Markov chains looks like they have converged and aren't drifting away from zero. In the next page we decided to take the mean and the median of

the 1000 iterations for each Markov chain for the element Mo.

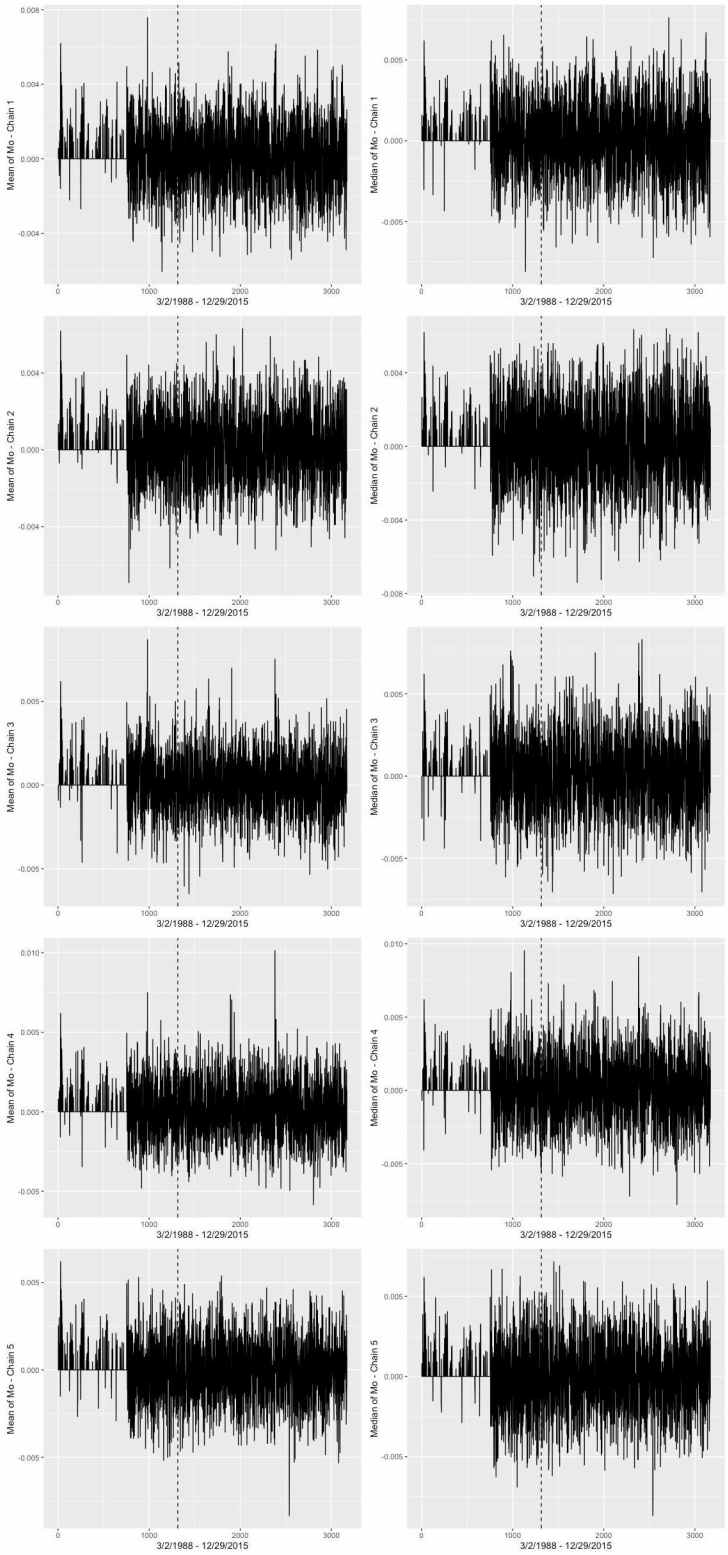

Figure 4.2: Mean and Median Plots of Mo for all 5 chains for each day

A dashed line was included in plots above of the mean and median of the imputed values of Mo. The dashed line indicates where Visibility dataset 1 ends and Visibility dataset 2 begins. From the plots above we can see that there are values that got plotted at zero prior to the dashed line. Since the change of algorithm that took into account values that were indistinguishable from zero did not happen until 2011 (after the dashed line), this meant that we were plotting zeros. Since we also had missing values prior to the dashed line, that meant that any values that were negative are from the imputed values.

The plots above shows the mean and median stayed relatively close and the five chains stayed consistent, which indicates that the model did well. Since the element Mo was missing in nearly half of the values in the dataset and the trace plots and plots of the mean and median this indicate that the model did well, we expect the same results for the other elements.

From the plots above, we can also see that the mean and median of imputed values dipped below zero. This is a downside when using the normal distribution to model the missing values because we have the possibility of having negative values. In order to prevent this issue, we thought about using a truncated normal distribution. We eventually decided against it and because it would greatly increase the computational effort (harder to program and slower to run) without any reason to believe the result would be affected in any meaningful way. After all, the positive values in the dataset will overwhelm the vague prior distribution on the mean concentrations, and result in positive numbers for the imputed values anyway. We also believed that by taking the mean and median of 1000 imputed values, we would not see any negative values unless it was influenced by the data, which we previously stated before included negative values.

5 Conclusion

In conclusion, it looks like the diagnosis that we ran indicates that the model did well. In order to get a better understanding if the model did well it we would be best if asked an expert in the field of atmospheric chemistry. Specifically, we could compare a known behavior or values of an element to the imputed values and see how well it did. Since we only looked at one element and have limited knowledge of atmospheric chemistry it is ultimately difficult to conclude how close the imputed values are to the actual real values.

5.1 Future Work

As previously stated, Hopke, Liu, and Rubin proposed three models and we only implemented two of the three. Due to time constraints we were not able to implement the third model, The Multivariate IMA Seasonal Time Series Model. We expect that being able to have a seasonal effect would ultimately result in better imputed values. The imputed values from model 2 and model 3 can also be compared to see which did better.

6 Reference

Air Quality Division (2012). Air Quality Monitoring at Denali National Park & Preserve, Alaska 2000 - 2003

Cleveland, W.S. and Liu, C. (1998). *Sum-difference time series model.* Technical Report, Bell Labs.

Dempster, A.P., Laird, N.M, and Rubin, D.B. (1977). Maximum likelihood estimations from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, series B 39, 1-38.*

Hopke, Philip K., Liu, Chuanhai, and Rubin, Donald B. (2001). Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic. Biometrics, vol. 57, no. 1, 2001, pp. 22-33., doi:10.1111/j.0006-341x.2001.00022.x.

R Core Team (2017), *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, Vienna Austria